TR-IT-0308

Problem Analysis and Strategies for Chinese Generation in TDMT System

宗 成庆　　山本 和英

Chengqing ZONG　　Kazuhide YAMAMOTO

1999 年 8 月

Abstract

Concerned the output of TDMT(Japanese-Chinese) system, the report analyzes the problems that the authors encountered initially in the Chinese generation, and then briefly summarizes the kinds of the problems. To achieve the goal of improving the accuracy and naturalness of Chinese generation, the report proposes repair strategies and presents the authors' opinions on proofreading the generation of TDMT system.

エイ・ティ・アール音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

Contents

1 Introduction

Although the strategies for target language generation are not the same in different machine translation systems based on different methods, it is the same goal, which is aimed at generating the correct and natural target language. However, it is a long way to reach the goal that human hopes. So, how to improve the accuracy and naturalness of the output is still hard work in research of machine translation(MT).

Concerned the output of TDMT(Japanese-Chinese) system, the report analyzes the problems that the author initially encountered in the Chinese generation, and then briefly summarizes the kinds of the problems. To achieve the goal of increasing the correct ratio and improve the naturalness of Chinese generation, the report proposes strategies for Chinese generation and presents the author's opinions on proofreading the generation of MT system.

2 Problem Analysis

2.1 The Problems in Chinese Generation of TDMT System

Having analyzed the output sentences of TDMT system, the author summarized the ill-formed sentences and classified them into 4 kinds as follows:

- The sentences are incomplete;
- There are redundant words or components in the generated results;
- The order of words or components in the output is confused;
- Some words are selected improperly and the sentence expresses obscure meanings.

Each kind of the ill-formed sentences mentioned above is illustrated respectively in remainder of the section.

2.1.1 The Incomplete Sentences

The incomplete sentences in generation include the following different cases:

1. The sentences lack subjects;
2. The sentences are short of predicates;
3. Several basic components(mainly including subject, predicate or object) are lacking in the same sentence;

4. The sentences are short of essential auxiliary words so that the naturalness of the sentence is too worse to understand.

It is true that some incomplete sentences are understandable and acceptable, especially in spoken dialog environment. In fact, some words of source sentences are often omitted by speaker. So, we have no reason to ask an MT system to add those words that are omitted originally in the source sentences. However, if the source sentence is not short of any components, and some words or components are never allowed leaving out in the specific dialog context, or some components are already indicated hiddenly in the source sentence, and if the result of generation is incomplete, the result will be hard to understand. See the following examples,

(1) *Input:* あした泊まることになってる鈴木直子といいますが
Output: 叫铃木直子。明天将住宿。

The sentence is the first one that a customer said to the hotel servicer. In this case, the subject of the Chinese sentence is generally not omitted, and it is better to generate the output sentence as a complete sentence. The proper result of the example should be: 我叫铃木直子, (我) 想明天住宿。

(2) *Input:* 子供にはツインそれから主人とわたしには和室をお願いしたいんですが
Output: 为孩子双人房, 要为我丈夫和我日本式房间。

In the example, the output should be consisted of two simple sentences. They are respectively: (1) 为孩子预定 (或 "准备") 双人房; and (2) 为我丈夫和我预定 (或 "准备") 日本式房间。 But in the real output, the first part consists of only a preposition phrase(为 (preposition) 孩子 (noun))and a noun(双人房). There is no predicate between preposition phrase "为孩子" and noun "双人房". So, the expression looks strange and it is hard to understand. The problem in the second part of the output is different. It belongs to the kind of word order confusion, which will be discussed later.

(3) *Input:* 予約をお願いいたします
Output: 要预定。

In fact the output is only a verb phrase, and it consists of an auxiliary verb "要" and a verb "预定". In the Chinese language it is possible that the speaker should only say a verb or a verb phrase to answer the question of his/her opposite side, but it is very strange that the speaker only say a verb phrase in his or her first sentence. At least the object of the verb phrase should appear if the agent is omitted. The proper expression of this sentence is: "(我) 要预定房间。".

(4) *Input:* でエキストラベッドのレンタル料は別に掛かるんですか
Output: 另外需要临时床的出租费?

According to the question mark at the end of the sentence, we can see that it is a query sentence. In fact the last character "か" in the source sentence states clearly that it is indeed a query sentence. But in the output sentence there is no mood auxiliary word

"吗", and there is not any other words to express the query tone. So, the generation result is not acceptable.

(5) *Input:* お願いしますコンチネンタルでお願いします
Output: 拜托了, 要大陆式。

The example seems no big problem just according to the form, but it is really ambiguous. The reason is that the sentence is short of an auxiliary word "的" after the word "大陆式". The word "大陆式" is an adjective(or it is treated as a distinction word in some system), and it can't be used as the object independently. The right expression should be made by the following modification: to add the word "的" after the word "大陆式" and to make the phrase "大陆式的" become a pronoun, which will play the role of object.

The incomplete sentences take a high proportion in the ill-formed output that I encountered. The ratio is about 60.4%. Although some of them are understandable and acceptable as I mentioned above, most of them still remain to be rectified further.

2.1.2 The Redundant Sentences

The redundant sentences also appear frequently in the generation results. In total 50 redundant sentences I found, 24(48%) sentences contain complete repeated words like the following examples (6)-(8). The rest of these 50 sentences contain superfluous words(like the following examples (9),(10) and (11)) or they are written in redundant expression forms(example (12) and example (13)).

(6) *Input:* はいこちらはフロントでございます
Output: 您好，这里是是总服务台。

(7) *Input:* お客様どうぞ御安心ください十六日の日は宵山ですよね
Output: 客人，请请放心。十六号号吧。好吗?

(8) *Input:* お部屋の状況を見てみますので少々お待ちください
Output: 看看看房间的情况，请稍等一下。

(9) *Input:* 来週の土曜日に予約をしたいんですけれども
Output: 想预订下星期星期六。

(10) *Input:* 三十分ですか料金はいくらぐらい掛かりますか
Output: 三十分吗? 需要大概多少钱费用?
In the Chinese language, the word "钱" has the same meanings as the word "费用", and generally they are not used after. The right generation of the example is: "三十分吗? 大概需要多少钱? " or "三十分吗? 大概需要多少费用? "

(11)*Input:* とってもおいしかったです

Output: 太很好吃了。

In the example, the adverbs "太" and "很" are repeated. In the Chinese language, except the words "特别" and "非常", a degree adverb is generally not allowed to follow another degree adverb. The proper result of the example is: 太好吃了。

(12)*Input:* お名前の名字は

Output: 你的姓名的姓呢?

The proper result corresponding to the input should be: 你的姓呢?

(13)*Input:* 失礼いたしましたはい五百十五号室のエイミーハリス様ですねお連れ様のお名前などお分かりになりますか

Output: 对不起，好，是五百一十五号房间的艾米·哈里斯小姐，对吗？知道您的同伴的姓名之类吗?

In result of the example, although the words "好", "的" and "之类" that are marked by underline in the output are indicated in the input sentence, the expression of the output is very strange. It doesn't follow the custom of the Chinese language. The proper result is: 对不起，(请问)是五百一十五号房间的艾米·哈里斯小姐，对吗？(您)知道您同伴的姓名吗?

2.1.3 The Sentences with Confusion of Word Order

The word order is often arranged improperly in the Chinese generation, especially the order of predicate verb. Generally speaking, in Japanese sentences the predicate verbs are used at the last part of the sentences, but the predicate verbs generally appear in the first part of Chinese sentences. So, whether the predicate verb is put in proper position will decide if the result of generation is expressed naturally. See the following examples:

(14)*Input:* そちらでウエアやシューズは貸し出しているんでしょうか

Output: 你们衣服和出借鞋吗?

The right result of the input sentence should be: 你们出借衣服和鞋吗?

(15)*Input:* ディズニーのアニメを借りたいと思うんですが今有りますでしょうか

Output: 准备想租的动画片，有现在吗?

The correct result of the input is: "准备租的(迪士尼)动画片，现在有吗？".

Besides the verbs, other words are also probably written in improper positions. As what you already saw in Example (10) given above, the position of adverb "大概" is changed in the correct result of generation.

2.1.4 The Sentences Expressing Obscure Meanings

In result of Chinese generation, some sentences express obscure meanings because the structures of sentences or some words are chosen improperly or even wrongly. In fact many sentences of source (Japanese) language can't be translated into target(Chinese) language sentences one to one. In many cases, it needs two or more Chinese sentences to express the meanings of one Japanese sentence exactly, it is the same verse while. Otherwise, the translated result may be obscure, and it will be very hard to understand. For example,

(16)*Input:* はいそれでも結構ですのでずっとよろしくお願いいたします

Output: 好, 那也没关系, 每天拜托了。

In output of the example, the phrase marked by underline is very hard to understand. In fact the meanings expressed in the input Japanese sentence may be expressed clearly by the following two Chinese sentences: (i) 每天都要给您添麻烦, (ii) 拜托了。

(17)*Input:* 無理を言ってすみませんじゃあそれでよろしくお願いいたします

Output: 对不起, 说难以办到的事情, 那么, 就这样, 拜托了。

It is the same with Example (16). The part with underline should be described by the following two simple sentences: (i) (我) 说了您办不到的事情, (ii) 让您为难了。

(18)*Input:* 申し訳ございません当ホテルでは一万七千円のお部屋からのご利用となっております

Output: 实在抱歉, 我们饭店是从一万七千日元的房间开始 使用 的。

In the example, the meanings of the source language sentence have been expressed in the target language sentence, but the expression form seems strange. It doesn't follow the expression custom of the Chinese language. The proper output is:"实在抱歉, 我们饭店是从一万七千日元的房间开始 租用 的。", in which the original word "使用" is changed into "租用". Another proper result may be: "实在抱歉, 我们饭店的房间是从一万七千日元开始起价的。". In this result not only the original word "使用" is changed, but also the position of three Chinese characters "的房间" are moved. However, it is difficult to judge whether an output sentence expresses ambiguously, especially to change the structure of output sentence.

The four kinds of errors and examples listed above only reflect the main problems in Chinese generation of TDMT system. Actually, in the output of generation, many trivial mistakes affect the naturalness of translated results, which can't be summarized in details. The author thinks that for the current MT systems, correctness is more important than naturalness.

2.2 Analysis on the Problems

The problems mentioned in the previous sub-section make the author ponder, "what are the reasons that cause the mistakes or worse naturalness in the target language generation ? " Now the reasons are summed up and analyzed in this section.

1. There is not enough knowledge on the target language to guide the generation of target sentence.

In research on machine translation, how to build effective knowledge base of the source language and how to design the algorithms and strategies for parsing source language sentences have been considered highly in the past decades, but how to use the knowledge on the target language to guide the generation is often overlooked or not treated as what is done to deal with the input of source language. The author thinks that the knowledge on target language is quite important in generation of translated results. In TDMT system, the sentence of the target language is generated from the structure which is the result of pattern transformation from the input to output. In fact, the structure of transformation maybe is wrong. See the following example.

(19)*Input:* どちらのタイプのお部屋でも静かさはそんなに変わらないと思います

The structure of transformation:

```
((L1(PRON 我)
    (VERB 想))
 (L2(NP(NP(NOUN 哪一种)
           (NP(AUX 的)
              (NOUN 类型)
              (NOUN 房间)))
    (L3((PP(PREP 关于)
           (NOUN 安静)
           (NOUN 这点))
        (VP(VERB 没有)
           (PRON 什么)
           (NOUN 区别))))
 ))
```

Output: 我想哪一种类型的房间，关于安静这点，没有什么区别。

The output sentence seems very strange. According to the grammar of the Chinese language, the structure of transformation is ill-formed. In the structure of trans-

formation, there are 3 levels: L_1, L_2 and L_3. The preposition phrase(PP) "关于安静这点" acts as the role of subject of L_3. L_3 is a clause(CS). But we hardly find proper syntactic rules of Chinese for generating the corresponding parsing tree. Here, L_3 may be parsed by the following rules:

NP + NP $\Longrightarrow$ NP
PREP + NP $\Longrightarrow$ PP
PRON + NP $\Longrightarrow$ NP
VP + NP $\Longrightarrow$ VP
PP + VP $\Longrightarrow$ CS

In the structure of transformation, L_2, which consists of NP and L_3, may construct a sentence(S) by using the following rule:

NP + CS $\Longrightarrow$ S

Thus, if L_2 is an independent sentence, L_1 will become superfluous. If L_2 is not an independent sentence, "PRON + VERB + NP + CS" is also a superfluous structure and it can't form a grammatical Chinese sentence.

According to the custom of the Chinese language, the output is unnatural and awkward. If a person expresses the sentence, PP is generally used at the beginning of a clause or a sentence and plays the role of adverbial respectively in the clause or in the main sentence.

The probable structure of transformation of the example may consist of 2 levels as follows:

```
((L1(PRON 我)
   (VERB 想))
 (L2((PP(PREP 关于)
        (NOUN 安静)
        (NOUN 这点))
     (NP(NOUN 哪一种)
        (NP(AUX 的)
           (NOUN 类型)
           (NOUN 房间)))
     (VP(VERB 没有)
        (PRON 什么)
        (NOUN 区别))))
 )
```

Output: 我想关于安静这点，哪一种类型的房间 (都) 没有什么区别。

In the structure, the PP("关于安静这点") plays the role of adverbial in clause L_2. The adverb "都" is a word that is expected to be added. The proper result may also be constructed as follows:

```
((PP(PREP 关于)
    (NOUN 安静)
    (NOUN 这点))
 (PRON 我)
 (VERB 想))
 (CS(NP(NOUN 哪一种)
       (NP(AUX 的)
          (NOUN 类型)
          (NOUN 房间)))
    (VP(ADVERB 都)
       (VERB 没有)
       (PRON 什么)
       (NOUN 区别))))
```

Output: 关于安静这点，我想哪一种类型的房间 (都) 没有什么区别。

In the result, the PP("关于安静这点") acts as the role of adverbial of the main sentence.

From the example we can see, besides checking the rationalization of generated structure, it is necessary to check whether some words in the generated result are superfluous, conflict or even wrong in semantic. It is also important to make the generated result follow the expression custom of target language.The author believes that it is beneficial to build a knowledge base on the target language and it is possible to guide the generation or to judge if the generation result is grammatical by using the knowledge of target language.

2. The one-to-one pattern from the source language to the target language can't deal with the source sentences with rich meanings.

In real corpus, some sentences of source language can be translated into target language one-to-one, but some of them can not be translated directly. It is hard to find a specific pattern of transformation corresponding to the source sentence or phrase, such as Example (16) and Example (17). In this case, $1:n$ ($n \in Interger$ and $n > 1$) pattern is needed. Here, n is a variable, and its value is different for different input.

However, it is difficult to select proper patterns in the mechanism of generation, but it is possible to generate the sentence of target language according to the semantic understanding results of the source sentence. If so, the structures of output sentences will not be limited by the source sentences.

3. The proofreading mechanism is needed to correct the results of generation.

Whatever happens, the improper results may exist. So, an MT system needs a special mechanism for proofreading the results of generation. However, the strategies for proofreading should be different from what is used in procedure of generation.

As it is analyzed above, the author thinks that the procedure of target language generation is an opposite one of source language analysis, and it is the same with parsing the source sentences, it is necessary to join multiple methods and strategies together in generation of target language.

3 Strategies for Chinese Generation and Proofreading

According to the ideas presented in the previous section, the author proposes strategies for generating and proofreading the output of Chinese sentences in this section. All strategies are built based on the knowledge bases.

3.1 The Knowledge Bases

3.1.1 The Syntactic Rule Base

The syntactic rules will be used in generation of the output sentences to check whether the structures of generated sentences are grammatical, or used in analysis of phrases, which is done in performing the related operations defined in the expression custom base. For description of the parsing rules, the part-of-speech of the Chinese language is classified. In our classification system, the part-of-speech includes noun(N), place name(W), direction word(F), time word(T), numeral(Q), classifier(L), pronoun(P), verb(V), auxiliary verb(X), judgment verb(J), adjective(A), adverb(D), preposition(R), conjunction(C), auxiliary word(H), mood auxiliary word(M), sound imitation word(Y) and idiom(I). The total number of the kinds of part-of-speech is 18. Division and expression of part-of-speech are described in detail in another report. The parsing rules are expressed as follows, and the main parsing rules for the Chinese language are described in the report titled as "The Part-of-speech and Parsing Rules for the Chinese

Language:

N + N $\Longrightarrow$ NP
NP(NHuman) + WP + "人"$\Longrightarrow$ CS
NP + C(Cwp/Ccw) + NP $\Longrightarrow$ NP
NP* + VP $\Longrightarrow$ CS

3.1.2 The Word Restraint Base

The word restraint base is designed to eliminate all probable repeated words in Chinese sentences, like "订购订购", "开始开始" or "星期星期" etc. In the Chinese language, some words may be used to express some special meanings in repeated forms, such as "非常非常 漂亮" or "考虑考虑 这个问题" etc. But some words are never allowed to write in repeated form, such as the words "开始" or "星期" etc. So, it is possible to build a base which records what kind of Chinese words maybe appear repeatedly and what kind of Chinese words are never allowed to repeat. The following shows a possible word restraint base designed by the author.

Format: #Pos: Expression;

where, "Pos" is a letter of part-of-speech of Chinese words, and the "Expression" is a symbol, a figure or an expression. Different "Expression" expresses different requirements. It is explained respectively as follows.

#N: 0; *In the Chinese language, all nouns are generally not allowed using in repeated form. If a noun is repeated, the repeated ones will be deleted. In the following items, it is the same with the meanings if the "Expression" is 0.*

#P: 0;

#Q: *; *It means the numeral is allowed appearing repeatedly any times.*

#V: -; *If a verb is used repeatedly, it is only allowed appearing once. e.g., "看看". If there is any extra repeated verb, the extra one will be deleted. In the following items, it is the same with the meanings if the "Expression" is "-".*

#X: -;

#A: -;

#D: $非常(Right:A), $特别(Right:A); *In all adverbs of Chinese, only two words "非常" and "特别" may be repeated when the right word is an adjective. If any other adverb is written in repeated form, or the conditions in the brackets after these two words are not satisfied, the repeated adverbs will be deleted.*

#C: 0;

#L: -;

#R: 0;
#H: 0;
#F: 0;
#Y: 3; *A sound imitation word is allowed appearing repeatedly 3 times at most. e.g., "砰砰砰". If a sound imitation word repeats more than 3 times, the extra ones will be deleted.*
#I: 0;
#J: 0;
#M: 0;
#T: 0;
#W: 0;

The "Expression" may be expanded by using defined logical expression.

3.1.3 The Expression Custom Base

In this base, some basic expression customs in the Chinese language are listed, which is used to find and eliminate some strange expressions in the generated results. The format of the records in the base is defined as follows:

#Conditions $\Longrightarrow$ RChoice;

Where, "Conditions" is expressed by symbols of part-of-speech or words, and the "RChoice" is result after modification operation. See the following examples,

N_1 和 N_2 和 $N_3 \Longrightarrow N_1$, N_2 和 N_3; *Here, N_1, N_2 and N_3 are all nouns.*
A_1 的 A_2 的 A_3 的 $\Longrightarrow A_1\ A_2\ A_3$ 的; *In this rule, $A_1, A_2 and A_3$ are all adjectives.*
$R_1\ R_2 \Longrightarrow R_1$; *Where, both R_1 and R_2 are prepositions.*
$H_1\ H_2 \Longrightarrow H_1$; *Where, both H_1 and H_2 are auxiliary words.*

Most of the probable rules in the expression custom base may be provided by Chinese linguists, and some of them may be obtained by statistical method. For exact description of conditions in the rules or in the syntactic rules, the semantic information of Chinese words may be involved.

3.2 The Strategies

The strategies for proofreading the generated sentences are based on the knowledge bases mentioned above. The basic procedure is described roughly as follows.

Input: A Chinese sentence or a phrase. Each word in the sentence or phrase is given the part-of-speech;

Output: The sentence or phrase that is rectified by the proofreader.

Assume all input sentences are generated under the guidance of the syntactic rules of Chinese. So, the structural nationalization of the input sentences will not be treated.

Procedure:

Step 1: The proofreader reads words one by one from the left to the right of input, and looks for any superfluous repeated word according to the word restraint base;

Step 2: The proofreader analyzes the result of Step 1 and checks whether some expressions in the result violate the customs of Chinese according to the expression custom base. If it violates, the related operations will be performed.

Step 3: The proofreader gives the result and then returns.

4 Discussion

In this section, the author proposes three topics for further research of Chinese generation in TDMT system, and each topic is discussed in detail as follows.

1. Research on strategies for generating sentences when some necessary patterns are absent.

The patterns are very important or even an only evidence to generate the sentences of target language. However it is impossible to find all patterns that cover all linguistic phenomena for the specific language pair(the source language and the target language). So, it is unavoidable that some necessary patterns are lacking for special sentences. In this case, how to generate a translated result according to the partial parsing results consequently becomes more meaningful.

2. Research on strategies for constitute combination and rearrangement.

Presently, in results of TDMT system, some translated sentences are often expressed by several separate constitutes like the following 2 examples:

(1)((J "京都名産品でお持ち帰りいただけるよう当ホテルでご用意してあるものです")

("京都，名产，可以带回。我们饭店，准备的。" . 6.7777843))

(2)((J “でそれで燃料は満タンにしてお返ししていただきますがよろしいですか”)
(“随后，燃料，退还，可以吗？”．5.928823))

These two sentences are not natural but their meanings are obviously shown. A Chinese speaker can find the respective result easily as follows,

(1') 我们饭店准备的京都名产，可以带回一些。

(2') 随后退还燃料，可以吗？

So, we have ground to think that it is very possible to rearrange the separate constitutes and to generate more natural sentences. It is very meaningful and beneficial to improve the naturalness of output in TDMT system.

3. Research on strategies for topic spotting and sentence reconstruction.

In some translated results of TDMT system, the order of words or phrases are confused, or the generated sentence is just a part of the result. The expressing meanings of the results are not very clear, but there is no big a problem for a Chinese speaker to understand what meanings the output sentences express. For instance,

Input: “それとできればですがね島などめぐってみたいと思いますんでそのへんが楽しめればいいなと思うんですが”

Output: “还有，最好。呢？因为我想想巡游看看岛子，所以该一带能享乐好呢？……想。”

The meanings of the output may be expressed: “因为(如果可能的话)我想巡游海岛，所以我关心是否能在这一带享乐好。”. As it is known to some people, in some special cases, a Chinese(or a Japanese) person can understand most of the meanings of a Japanese(or Chinese) sentence just according to some Chinese characters. This phenomenon has the author think that maybe there is a way to pick out the topic of the input sentence according to the parsing results, and the output sentence may be rebuilt according to the topic and some words.

Any how, research on every topic mentioned in the section will involve many problems of natural language processing, including syntactical analysis, semantic analysis and context analysis in the Chinese language. However, to combine the different strategies and methods must be beneficial and practicable.

5 Acknowledgments

During the period when the author worked in Department 3 of Interpreting Telecommunications Research Laboratories of ATR, Mr.Satoshi Shirai, head of Dept.3, senior researcher Eiichiro Sumita and Ms.Yoshiko Kakitani had beneficial discussions with the author on the problems that the report involves and helped the author a lot in many other aspects. The author would like to thank all of them for their generous help, and the author is grateful to researcher Yves Lepage, Setsuo Yamada, Michael Paul, Ms.Chiaki Morita and Mr.David Brown for their generous help. The author also would like to say a very big thank to all of those, whose names are not mentioned above, who lent the author a hand in any aspects as well.